Indian Journal of Science and Technology, *Vol 8(27), DOI: 10.17485/ijst/2015/v8i27/60811, October 2015*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Novel Approach for Computing Semantic Relatedness of Geographic Terms

Mozhdeh Nazari Soleimandarabi[1]* and Seyed Abolghasem Mirroshandel[2]

[1]Department of Computer Engineering, Rasht Branch, Islamic Azad University, Rasht, Iran; mozhdeh_nazary@yahoo.com
[2]Department of Computer Engineering, University of Guilan, Rasht, Iran; mirroshandel@guilan.ac.ir

Abstract

Background: Computing semantic relatedness measures are extensively employed in the field of Natural Language Processing (NLP) and play pivotal role in Geographic Information Science (GIS). **Methods/Analysis:** Noteworthy, despite the significance of semantic relatedness in geographic domain, its role has been almost ignored. While most of the proposed measures in this context are only able to compute semantic similarity, in this paper, the notion of geo-semantic relatedness is discussed and from which a novel approach for computing semantic relatedness of geographic terms is proposed. The proposed method utilizes term's definition from geographic lexicon. Whereas lexical definition demarcates the boundaries of a term and provides valuable semantic space for deducting the meaning of a term, it can have prominent impact in the efficiency of the proposed approach. Furthermore, this approach exploits Wikipedia as semantic resource that has considerable performance in application of semantic relatedness. **Finding:** The cognitive plausibility of the proposed approach is evaluated on GeReSiD dataset. Compared to the previous state of the arts, using proposed approach results significant improvement in correlation of computed relatedness score with human judgment to 0.73. **Conclusion:** Additionally, the proposed approach not only prospers higher perception and adoption, but also it has greater applicability in real world problems and is confronted with fewer limitations. Moreover, the proposed method can perform disambiguation in geographic domain properly.

Keywords: Definition, Geographic Lexicon, Geo-Semantic, Geographic Semantic Relatedness, Wikipedia

1. Introduction

Computing semantic relatedness is expressed as a fundamental task in natural language processing and its goal is to identify a measure that can express the strength of semantic association among a pair of concepts including classical and non-classical relations .On the other hand, according to this large amount of information that researchers in both academic and professional community are confronted, the necessity to extract concept from the chaotic repository of implicit concepts is felt more than ever and users expect to retrieve the most relevant documents to their queries[2]. By considering this issue, the role of semantic relatedness measure is significantly highlighted. Noteworthy, whereas up to 80 percent of human's decisions are related to a space or locations and geographic knowledge is a crucial benefit in human activities, semantic web technologies attempt to provide a platform to share geographic knowledge[2–4]. As an example consider that people are doing a research about Berlin, Bill Gates or any other concepts and want to know the places that are related to these concepts and the reason of their relations. Consequently, geographic semantic relatedness is an emerging issue, which has wide application in discovering terms in geographic map[5], geo-information retrieval[6] and geographic information science[7].

Semantically related term are connected via any kinds of relations, while semantic similarity is identified as particular subset of semantic relatedness involving hyponym-hypernym relations among terms[1,8]. For example,

*Author for correspondence

"lake" and "pond" are semantically similar while "lake" and "fish" are not similar but they are semantically related. Additionally, geo-semantic relatedness is defined as a specific subset of semantic relatedness focusing on relations based on geographic dimensions. It means that terms are geographically related to the degree to which they belong to geographic domain[3].

Measuring semantic relatedness requires specific background knowledge about terms and concepts[9,10]. Over the last decades, a variety of measures have been developed for computing semantic relatedness of terms and the importance of this issue has been also favoured in geographic domain[3]. Various methods considering their employed background knowledge are divided into two distinct types of knowledge-based and corpus-based. Knowledge-based approaches rely on semantic relations of concepts in ontology or taxonomy such as Word Net. In contrast, corpus-based approaches do not require explicit relations among concepts and can compute semantic relatedness among terms based on their co-occurrence in large corpus of documents[1].

There is fundamental trade-off among knowledge-based and corpus-based approaches. Knowledge-based approaches require expert-authored background knowledge. Notably, constructing and maintaining such this kind of knowledge bases in a specialized domain is time consuming and costly. Moreover, these knowledge bases are limited to a special domain and do not cover wide range of concepts. On the other hand, corpus-based approaches are not confronted with any specific limitations and they tend to cover wider set of concepts[11].

In previous works, it has been revealed that the performance of corpus-based measures is significantly superior to knowledge-based measures[1,9,12]. Notably, knowledge-based measures have been slightly employed in the field of geographic information science, whereas applications of corpus-based measure despite their high performance have been almost ignored in this filed.

On the whole, to the best of our knowledge, most of measures presented in geographic domain employ relations in a particular expert-made knowledge base and are only able to compute semantic similarity and do not consider all relations among terms[3]. To fill this lacuna, this paper presents a corpus-based measure to compute semantic relatedness of geographic terms based on their lexical definition extracted from geographic lexicon and employing Wikipedia as background knowledge.

Whereas concepts in geographic domain are highly fragmented and users of this field have limited background knowledge about specialized terms and concepts, geographic domain specific lexicon plays a pivotal role across this diverse information. In other words, geographic lexicon indicates the usage of a term in relevant scope and specifies its applications. Therefore, ambiguities about the meaning of a term are entirely eliminated and reasoning about the relatedness among concepts can be performed more precisely.

The reminder of this article is organized as follows: Related works in the area of both general and geographic domain are presented in section 2. The proposed measure for computing semantic relatedness of geographic terms is outlined in section 3 extensively. Empirical evaluations and experimental set up are indicated in section 4. Conclusions and directions for future research are also presented in Section 5.

2. Review of Literature

Due to geographic domain, computing relatedness is an important technique for discovering functional relation among places and constructing lexical resources. Over the years, large number of semantic relatedness measures have been conducted with special respect to geographic domain[13]. Various measures leverage lexical and semantic information of terms and concepts usually encoded in some background knowledge to be able to judge about the meaning of the terms. Accordingly, previous works considering their employed background knowledge are classified in to two types: knowledge-based and corpus-based which are extensively discussed in the following. It must be noted that whereas the focus of this paper is on geographic domain, the related work section contains researches conducted in both general and geographic domain.

2.1 Knowledge-Based Models

Knowledge based techniques rely on semantic relations among terms in ontology, taxonomy or semantic network of background knowledge bases for computing semantic relatedness. In other words, most of the measures presented in this field require expert-made background knowledge about terms and concept that encodes the relations among them and relatedness is computed by considering the inverse taxonomical distance between two concepts. Some

methods especially earlier ones have leveraged dictionaries and thesaurus[1]. Moreover, Word Net is a well-known resource that encodes various relations among terms and has been extensively used as background knowledge in task of computing semantic relatedness[14].

The primary methods that employed Word Net as background knowledge determined the relatedness among terms based on their distance or the length of the path connecting them in ontology[15]. The length is computed by counting the number of nodes in a path. The shorter path indicates higher relatedness among concepts. Although these measures were simple and performed fairly well, they were confronted with some limitations. Whereas these measures employed taxonomic links, they considered similarity rather than relatedness. Moreover, they could not consider that the higher concepts in taxonomy are more abstract. In order to overcome these limitations, some methods have been developed by[16,17] which employed the notion of the Lowest Common Subsumer (LCS) of two concepts in taxonomy. LCS is the most specific concepts shared from the leaf to the root of hierarchically. Additionally[18] suggested limiting the length of the path by considering direction changes. Based on their hypothesis, changing directions and long path correlate negatively with semantic relatedness.

Other Word Net-based semantic relatedness measures depend on Information Content (IC). Based on this criteria, the relatedness among a pair of terms refers to the amount of information that the share. Following the similar line of research[19], proposed a measure which associates probability to each concepts from statistics in a large corpus of texts and computes IC among two concepts with respect to their LCS. Furthermore[20,21], augment the IC of two concepts with the sum of the information content of individual concepts.

Whereas dictionaries such as Word Net contain short gloss for each concept explaining the meaning of corresponding concept, some measures have been proposed to compute relatedness among concepts using information provided by glosses[22]. In other words, they consider the amount of term overlap in the glosses of two concepts. The higher overlap means higher relatedness. Considering the fact that Word Net glosses are short and are not able to provide extensive information about concepts, proposed[23] a measure which creates co-occurrence matrix of corpus made by Word Net glosses. Therefore, each concept has associated context vector and relatedness is computed by determining the cosine among two corresponding gloss vector.

It must be taken into consideration that building and maintaining lexical resources such as Word Net is time consuming and expensive and their coverage is also limited in dealing with domain specific technical terms[1]. To fill this lacuna, recent researches have focused on collaboratively built lexical resources such as Wikipedia as background knowledge base. In the following[24], proposed a measure which leverage Wikipedia category graph for computing semantic relatedness. Particularly, they applied Word Net path techniques on Wikipedia article graph. Moreover[25,26], proposed a directed graph using internal links of Wikipedia for computing relatedness.

Noteworthy, various methods have employed knowledge-based techniques for computing semantic relatedness in geographic domain. Matching Distance Similarity Measure (MDSM), proposed by[27], was one of the first semantic relatedness measures, which has been developed specifically for geographic domain. According to this method, asymmetric values for relatedness of spatial entity classes were obtained based on their degree of generalization within a hierarchical structure. In the other word, it compared entity classes in terms of their distances in the semantic structure that was defined by the semantic relations.

Furthermore[2], developed a method which used Wikipedia article graph for computing semantic relatedness. Based on their notion, the relatedness score was computed by assigning weight to spatial referred articles in Wikipedia article graph. It is worth noting that semantic networks which encode knowledge and meanings in the form of graphs, have been also used in computing semantic relatedness of geographic terms. Furthermore[28], developed a method which was based on some forms of structural distance between nodes (e.g. edge counting) or on the topological comparison of sub graphs.

Recently[29], developed a method which leveraged Volunteered Geographic Information (VGI) for computing semantic relatedness. VGI is a large reusable unit of geographic knowledge generated by heterogeneous information communities. Using VGI information, they applied graph-based measures of semantic relatedness on Open Street Map (OSM) semantic network. Whereas, OSM semantic network consists of noisy and ambiguous data, in similar work they enriched the OSM semantic model with semantic web resources[30].

2.2 Corpus- Based Models

Unlike knowledge-based models, corpus-based models do not require structured resources as background knowledge and explicit semantic relations among term. In other words, corpus-based measures employ statistical analysis on background corpus to compute semantic relatedness[9]. As a result, background knowledge is collected at the level of terms rather than concepts. These measures rely on this hypothesis that related terms are occurred in similar context[8]. Notably, unlike knowledge-based measures that have been extensively used in geographic domain[31], the notion of corpus-based measure unlike their superior applicably have been almost ignored with some exceptions[2,32].

Going beyond simple co-occurrence, Latent Semantic Analysis (LSA)[33] is one of the most prominent measure which considers pattern of co-occurrence in individual sentence rather than the total number of co-occurrences . Moreover, LSA does not rely on human-organized knowledge and it is a dimensional reduction technique, which applies Singular Value Decomposition (SVD) on term-documents matrix to identify the most important dimensions in data. In order to improve the efficiency of LSA[34], proposed a measure which used Non-Negative Matrix Factorization for reducing the dimensions of term-document matrix and used various global and local weighting method for constructing term-topic matrix[35].

Explicit Semantic Analysis (ESA) is another approach which employs distribution of terms in a corpus of unannotated natural language text[36]. Based on ESA, each term is presented as a vector of Wikipedia concepts and semantic relatedness is computed by comparing vectors of a pair of terms in multidimensional vector space. Although ESA presents high correlation with human judgment, it is also confronted with some limitations. The obvious drawback of this approach is that its concepts might be difficult to interpret in natural language and polysemy is also considered as serious problem. To overcome these problems, some measures have been proposed[37,38].

As it was previously mentioned, corpus-based models have been rarely explored in GI Science. More recently, by considering spatial co-occurrences features[39], extracted a relatedness measure directly from Open Street Maps (OSM) vector data. Moreover[32], proposed a method for computing relatedness of geographic terms which leveraged frequently used semantic measures for computing relatedness such as ESA[36] and generated human readable explanation by mining text hyperlinks and Wikipedia category graph. Furthermore[11], proposed a hybrid method to quantify semantic relatedness of lexical definition. Based on their idea, related terms tend to be defined using similar terms. This measure combined existing Word Net and paraphrase detection techniques for computing semantic relatedness.

3. Computing Semantic Relatedness of Geographic Terms based on Lexical Definitions

This section outlines an approach to compute semantic relatedness of geographic terms using definitions extracted from geographic lexicon. Moreover, the proposed method leverages Wikipedia as the source of semantic relatedness. The purpose of this measure is to quantify the semantic relatedness of two given geographic terms $Term_a$ and $Term_b$ as a real number, based on their lexical definitions Def*Term a* and Def*Term b*. The intuition behind our approach to compute geo-semantic relatedness is that similar terms exist in definitions of related terms (i.e., if a pair of terms are related, their definitions consist similar terms). The infinite regression that would ensue is avoided by using Wikipedia to compute the relatedness scores of definitional terms. In fact, in order to eliminate the dependencies of the proposed method to lexical definitions, Wikipedia is used for weighting. This criteria stands on that related terms co-occur in the same articles. The four steps of the relatedness algorithm are the following:

1. Extracting definitions of $Term_a$ and $Term_b$ from geographic lexicon.
2. Pre-processing the definitions in order to extract the best descriptors.
3. Constructing a semantic interpreter for weighting the descriptors and creating a vector for each descriptor based on Wikipedia articles.
4. Summing descriptor vectors to create a single vector for each term and computing relatedness score by comparing the vector of each term.

The architecture of the proposed approach is presented in Figure 1. For illustrative purposes, lexical definitions from the OSM Semantic Network were considered. The remainder of this section describes in detail the four steps to compute the semantic relatedness of geographic terms.

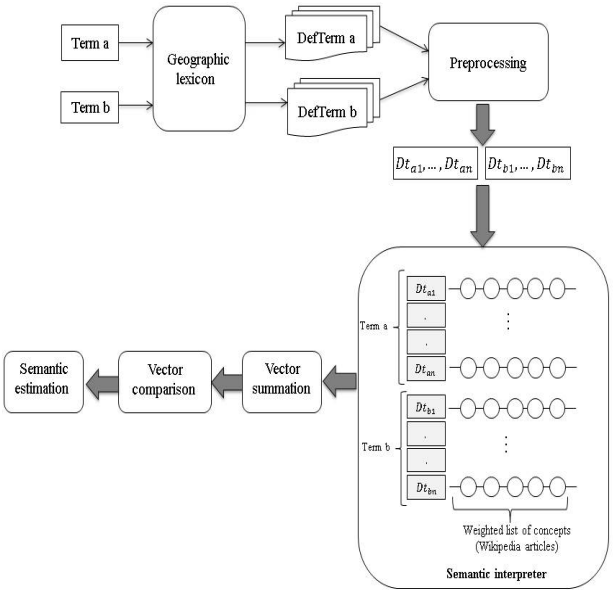

Figure 1. Architecture of the proposed approach for computing semantic relatedness.

3.1 Extracting Definitions of Terms

The lexical definition of a term presents the meaning of a term in common usage and plays a pivotal role in creating a shared semantic ground. Whereas, the central point of this paper is on geographic domain, a particular geographic lexicon is required to present the properties of concepts in this scope. Lexical definition of geographic terms is specifically focused on the context of Volunteered Geographic Information (VGI)[7], which is initiated by heterogeneous information community. Based on VGI project, a large number of humans co-operate to produce reusable units of geographic knowledge. It must be noted that geographic lexicon is able to both describe the properties of a concept and manifest the usage of a term based on the desired scope. Furthermore, geographic lexicon can provide extra information about terms in corresponding domain, which can help user to interpret the meaning of a term efficiently. Accordingly, geographic lexicon can provide rich background knowledge for computing semantic relatedness of geographic terms[11].

Given a geographic $term_a$, Def $term_a$ is its definition containing a set of terms and punctuation $\{t_{a1} \dots t_{an}\}$which contribute to determine the overall meaning of $term_a$. In order to provide an adequate definition, OSM wiki website, which hosts numerous lexical definitions of geographic term is leveraged[29].

3.2 Pre-Processing the Definitions

As mentioned in the previous section, the lexical definition of a geographic term is a string consisting definitional term which cooperates to conduct the meaning of geographic $Term_a$ It is obvious that all these terms are not good descriptors for conveying the meaning of a term, therefore preprocessing is an anobligatory step, which is yield to facilitate the use of them. Preprocessing consists of three main stages: Tokenization, removing stop words and stemming. The specified goal of preprocessing step is eliminating the noise and identifying the most suitable descriptors. Based on the proposed method, the suitable descriptor is able to convey the meaning of a term and determine the differences among various terms.

First stage of pre-processing is tokenization where each geographic definition is divided into separate tokens. In other words, this phase of process consists of converting lexical definitions to an array of its terms. After tokenizing, the less meaningful terms known as stop words must be removed. Therefore, the remaining terms are appropriate modifiers that can indicate the exact meaning of a term. The last stage is stemming to convert the terms into root form. Stemming is the process for reducing derived terms to their stem. Porter Stemmer was leveraged in the proposed approach[40]. At the end of this stage, a geographic term is presented as an array whose components are definitional terms.

3.3 Constructing a Semantic Interpreter

According to previous step, lexical definitions were transformed into definitional terms where each term is a valuable descriptor of its corresponding geographic terms. Moreover, this presentation does not take into consideration syntactical structure. In order to reduce dependencies of the proposed method, terms are entered to semantic interpreter. While descriptors are given to semantic interpreter, it rates all of Wikipedia article based on their relevance to input descriptors. In other words, semantic interpreter construct a term-document matrix for each geographic term where its columns are Wikipedia articles and its rows are definitional terms extracted from preprocessing step. Subsequently, each elements of this matrix represented the weight of a term in a Wikipedia article. The main goal of weighting is enhancing the precision and eliminating the dependencies of the proposed measure to lexical definitions.

As it was previously mentioned, the hypothesis of the proposed method states that related terms are used in the same article and weighting is used to show the influence of each definitional term in a specific article. A popular approach term frequency inverse document frequency TF-IDF is used for weighting. IDF score represents how a common term is in the whole corpus and TF is the number of occurrence of a specified term in a document[8].

In other words, if $\{t_{a1} \ldots t_{an}\}$is a set of definitional term of geographic $Term_a$, $< v_{a1} \ldots v_{an} >$are semantic vectors for each descriptor constructed by semantic interpreter. Each of these vectors represents the weight of a specific definitional term in Wikipedia articles. In particular,w_j presents the intensity of semantic relations among descriptor t_{aj} and Wikipedia article d_j, where $\{d_j \in d_1, \ldots, d_n\}$ and n is the number of Wikipedia articles. Therefore, the semantic vector for $term_a$ is a vector with length of n where its elements represent the weight of term t_{a1} in document d_j. The structure of semantic interpreter is illustrated in Figure 2.

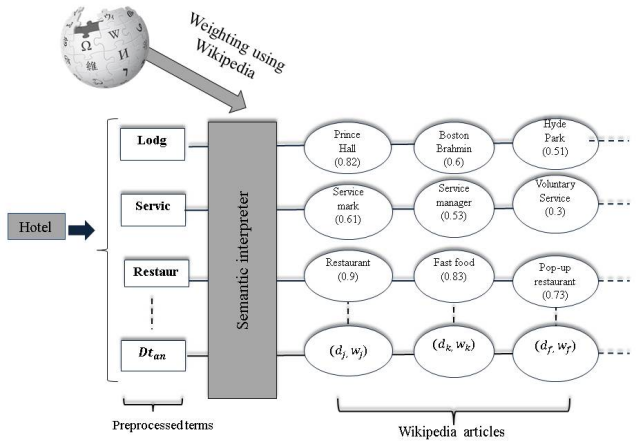

Figure 2. Creating weighted vectors using Wikipedia for descriptors of term "Hotel".

3.4 Vector Summation and Comparison

In principle, the relatedness between two geographic terms is computed using simple vector similarity measure as the inverse distance between vectors. Therefore, descriptor's vectors for extracted from previous step which have equal dimensions (Wikipedia articles) must be summed and create a single vector presenting the meaning of $Term_a$. Consequently, the single vectors for $Term_a$ and $Term_b$ are created as follows:

$$\{V_{Ta} = v_{a1} + v_{a2} \ldots + v_{an}\} \quad (1)$$

$$\{V_{Tb} = v_{b1} + v_{b2} \ldots + v_{bn}\} \quad (2)$$

Where V_{Ta} represents the semantic vector for $Term_a$ and V_{Tb} represents the semantic vector for $Term_b$which are initiated by summation of their descriptor's vectors[8]. Finally, created semantic vectors are mapped into multi-dimensional space of Wikipedia articles and compared using cosine similarity measure as follows:

$$\cos(x, y) = \frac{V_{Ta} \cdot V_{Tb}}{|V_{Ta}||V_{Tb}|} = \frac{\sum_{i=1}^{n} V_{Tai} V_{Tbj}}{\sqrt{\sum_{i=1}^{n} V_{Tai}^{2}} \sqrt{\sum_{i=1}^{n} V_{Tbi}^{2}}} \quad (3)$$

4. Experiments

Empirical evaluations of the proposed method for computing semantic relatedness of geographic terms are presented in this section. The aim of these evaluations is to reveal the priorities of the proposed method in comparison to other well-known existing knowledge-based and corpus-based measures in geographic domain. Moreover, evaluation is a prominent factor in presenting the strengths and weaknesses of various semantic relatedness measures.

Two complementary approaches have been typically utilized for evaluations of semantic relatedness measures: In-vivo and In-vitro. Due to in-vivo experiments, the effectiveness of a semantic relatedness measure is qualified using specific application. In other words, the relatedness measure is applied to a specific task such as word sense disambiguation or information retrieval. Therefore, the performance of the relatedness measure is revealed by considering how satisfactory the task is performed. In contrast, In-vitro evaluation is done by comparing human judgments on a set of pairs of terms to machine generated results on the same benchmark dataset. It means that based on this method, a set of pairs of terms is given to humans and they estimate the relatedness among terms in certain scale. The dataset is then given to machine and the correlation coefficient among human judgments and machine-generated results is computed using Spearman and Pearson correlation coefficient.

Whereas in-vivo evaluation depends on specific framework and employed background knowledge and it also entails influence of specific application parameters,

its result may not be entirely precise and accurate. Consequently, in-vitro evaluation, which is independent of background knowledge and other parameters, is leveraged in our experiments. In addition, based on in-vitro evaluation method, various semantic measures are compared at the same situation without any regards to their background knowledge.

4.1 Benchmark Dataset

A set of pairs of geographic terms, the GeReSiD[3] benchmark dataset was employed as gold standard in our experiments. This dataset includes 97 geographic terms combined into 50 phrase pairs. The degree of relatedness among each pair is originally collected from 203 native English speakers through an online survey. This dataset is larger than any other dataset in geographic domain and it is certainly the only dataset that considers both relatedness and similarity and outlines differences among them[3]. GeReSiD is available online and it can be used as a valuable resource for evaluating geographic models and defining correlation of experimental results with human judgment in GI Science.

Additionally, to increase the usability and clarity of this dataset, its terms have been mapped to their corresponding synset in Word Net and OSM. Moreover, this dataset covers terms from natural entities to human-made features and its term conveys uniform distribution. It means that the number of high related, middle related and low related pairs comply the same distribution. In addition, the differences among semantic relatedness and similarity is clearly specified in this dataset. It is due to this issue that the score of semantic similarity is generally lower than the semantic relatedness score among terms and it clearly reflects this fact that semantic similarity is specific kind of semantic relatedness[3]. According to the mentioned advantages and sates of being open-source, this dataset has been employed in our experiments. It must be mentioned that GeReSid is the only dataset that captures differences among geo-semantic relatedness and similarity in geographic domain, while other existing data sets only consider semantic similarity.

4.2 Experimental Set Up

As previously mentioned, the impact of relatedness measures have been almost ignored in geographic domain and majority of works dedicated in this field can only consider semantic similarity among terms. Therefore, the main issue of this paper is presenting a novel approach that is able to compute semantic relatedness of geographic terms. In order to show the superiorities of the proposed method, it has been compared to a wide range of knowledge-based measures, which are extensively employed in geographic domain and some state-of-the-art corpus-based measures that has high correlation with human judgment generally.

The experiments of this paper are divided into three groups. In first set of experiments, the semantic relatedness of 50 pairs in GeReSiD dataset was computed using 10 Word Net-based measures. These sets of experiments were conducted to support this claim that the existing knowledge-bases measures do not present high precision in computing semantic relatedness of geographic terms. Accordingly, the semantic relatedness among a pair of terms is computed using different aspect of Word Net taxonomy. Wordnet: Similarity package was employed for this set of experiments[41].

The second set of experiments focused on implementing some corpus-based measure in order to compare their performance to the proposed method. To claim this issue, Latent Semantic Analysis (LSA)[33] and Explicit Semantic Analysis (ESA)[36], the state-of-the-art semantic relatedness measure, were executed on GeReSiD dataset. Esalib[36] package was employed for implementing ESA and Gensim[42] package was used for implementing LSA. The aim of this set of experiments is revealing the precise of corpus-based measures in comparison to knowledge-based measures and presenting the superiorities of the proposed measure. For implementing ESA and LSA approaches, early spring 2013 version of Wikipedia, containing about 4 million articles was used.

The third set of experiment is conducted to implement the proposed measure. Whereas the proposed measure is corpus-based measure which uses Wikipedia as background knowledge, the same Wikipedia corpus was employed. Noteworthy, Wikipedia contains large amount of noisy information that are not interpretable by machines and it must be pre-processed. At first, Wikipedia XML dump was parsed. In the following small and overly specific concepts were removed (those having fewer than five incoming or outgoing links)[43]. The texts were then processed by removing the stop word and stemming the remaining terms. Porter stemmer was employed for this task. The remained distinct terms were served for implementing the proposed method.

As previously mentioned, the proposed method requires lexical definitions. Lexical definitions of 97 terms were extracted from OSM network for applying the experiments of this paper. The extracted definitions were pre processed in order to extract the valuable descriptors and decrease the complexity of computation. The average lengths of lexical definitions were 46 terms and only small numbers of definitions were longer.

4.3 Empirical Results

In this section, extensive experiments for evaluating the efficiency of the proposed method and comparing its performance to the other state of the arts in application of computing semantic relatedness of geographic terms have been conducted. As previously mentioned, to better evaluate the proposed method, a set of knowledge-based and corpus-based measures were also implemented. The correlation between empirical results and scores determined by human judgments on GeReSiD[3] dataset is presented in Table1. This table reflects the results and accuracy of applying our methodology for estimating the relatedness of geographic terms in comparison to other well-known measures.

Table 1. The correlations among different algorithms and human judgments based on Spearman (ρ) correlation coefficient on GeReSiD dataset.

Algorithm	Spearman's Correlation (ρ)	p-value
Proposed method	**0.73**	0.0017
LSA[33]	0.71	0
ESA[36]	0.68	0.0032
HSO[18]	0.41	0.0033
Lesk[44]	0.39	0.005
Vector[23]	0.56	0
Resnik[19]	0.26	0.0739
Lin[20]	0.39	0.0056
Jcn[21]	0.31	0.0266
Lch[17]	0.37	0.0087
Wup[16]	0.33	0.0183
Path[41]	0.45	0.001

The values shown in Table 1 represent Spearman correlation (ρ) among human judgments and scores produced by various measures[1,3,36]. As it is clear, the proposed measure yield substantial improvement (ρ = 0.73) over the most prominent knowledge-based measures, which are extensively used in geographic domain. Notably, the proposed measure also achieves much better results in geographic domain than LSA and ESA methods, which leverage Wikipedia as background knowledge. Therefore, it can be said that the proposed method can perform significantly better in application of geo-semantic relatedness.

Although the proposed measure reflects higher correlation with human judgment, it is also confronted with some drawbacks. Figure 3 presents the histogram of value distribution of the proposed method in comparison to the human judgment of GeReSiD dataset[3]. As it is clear, the value distribution determined by human judgment comply uniform distribution. In contrast, the histogram of the value distribution estimated by the proposed method has skewness. It confirms that unlike high correlation of the proposed method with human judgment, it also has some drawbacks in identifying the exact value of relatedness between two pairs of terms. This issue can play important role in future development of the proposed method.

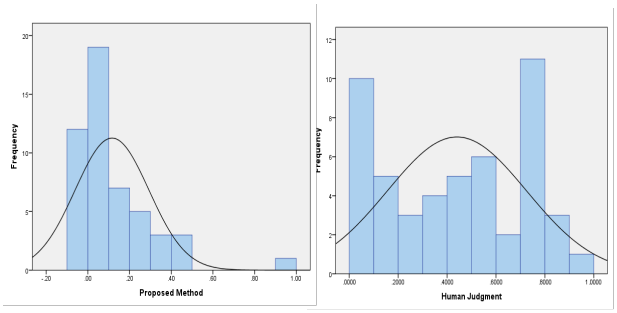

Figure 3. The histogram of value distribution of the proposed method and human judgment on GeReSiD dataset.

Additionally, to illustrate the strength of the proposed method in geographic domain, the extracted ten highest scoring concepts (Wikipedia article) for two terms of hotel and motel are presented in Table 2. These concepts are extracted when concepts of each term in each vector are sorted in decreasing order of their scores (the weight of each term in each Wikipedia article). The top ten concepts are the most relevant ones to the input terms. As it is obvious, our proposed methodology is capable of performing word sense disambiguation in geographic domain. In other words, all extracted concepts for these two terms are spatial article of Wikipedia, which shows

the strength of the proposed method in confronting to geographic terms. It means that the proposed method can interpret the meaning of terms according to geographic domain, while the other corpus-based measures consider general article without paying any particular regards to geographic domain.

Table 2. First ten concepts in sample interpretation vectors of terms "Hotel" and "Motel".

Input: Hotel		Input: Motel	
1	Hotel	1	Motel
2	Fairmont Hotels and Resorts	2	Hotel
3	At Bertram's Hotel	3	Roach Motel
4	List of hotel chains	4	Hits Plus
5	Benson Hotel	5	The Motels
6	King David Hotel	6	Super 8 Motels
7	Brand	7	Parking lot
8	Chicago, Illinois	8	United states
9	Hotel rating	9	Highway
10	Joan Crawford	10	Railway stations

5. Conclusion and Future Work

In this paper, a computational method for semantic relatedness of geographic terms based on their lexical definition is proposed. Based on previous studies, most of the works dedicated in geographic domain for computing semantic relatedness employs semantic relations of concepts in taxonomy and ontology[3,13]. Whereas corpus-based measures have shown higher performance and accuracy in computing semantic relatedness in previous studies[1,9,36] and the notion of these measures has been entirely unexplored in geographic domain, we tackled a challenge of devising a corpus-based measure for computing geo-semantic relatedness. This measure combines lexical definitions of terms extracted from geographic lexicon with methodologies of corpus-based measure.

Empirical evaluation confirms that the proposed method leads to substantial improvement in computing semantic relatedness of geographic terms in computing semantic relatedness of geographic terms and phrases. Compared to previous knowledge-based methods[14], using the proposed method results notable improvement. Moreover, unlike knowledge-based measure, the proposed method is not confronted with particular restriction and can be easily applied in real world application. Furthermore, empirical results have also shown the higher accuracy of the proposed method in comparison to the state-of-the-art corpus-based measures. Although the implemented corpus-based measures (ESA[36] and LSA[33]) have high precision in general domain, the results revealed the issue that the proposed measure is significantly superior to them in geographic domain.

It must be noted that the proposed approach provides a highly plausible measure of semantic relatedness of geographic terms. In other words, whereas this approach employs geographic lexical definition and natural concepts as background knowledge, it can eliminate ambiguities about the meaning of a term and it is easy to explain to human users.

The proposed measure of this paper can be applied to a number of natural language processing tasks using the same or any other lexicon. It means that this measure can be applied in another domain such as biomedical using a particular lexicon on that field. Moreover, this measure can be used in application of explanatory search for discovering relations among terms in geographic map. Additionally, it can be employed as a fundamental task to perform query expansion in geographic information science. Another usage of this measure is in data mining for clustering related terms and finding a special pattern among them.

Possible extensions to this work focus on using geographic specific domain background knowledge instead of Wikipedia and employing a richer geographic lexicon for extracting lexical definition of geographic terms. Moreover, another datasets can be employed for revealing the strengths and weaknesses of the proposed measure. Consequently, using this measure and methodologies behind it can leads to wide range of future research and valuable applications of semantic relatedness in geographic information science.

6. Reference

1. Zhang Z, Gentile AL, Ciravegna F. Recent advances in methods of lexical semantic relatedness – a survey. Natural Language Engineering. 2013; 19:411–79.
2. Hecht B, Raubal M. GeoSR: Geographically explore semantic relations in world knowledge, in The European Information Society. 2008. p. 95–113.
3. Ballatore A, Bertolotto M, Wilson DC. An evaluative baseline for geo-semantic relatedness and similarity. GeoInformatica. 2014. p. 1–21.

4. Khusro S et al. Linked open data: Towards the realization of semantic web-a review. Indian Journal of Science and Technology. 2014; **7**(6):745–64.
5. Haklay M, Weber P. Openstreetmap: User-generated street maps. Pervasive Computing, IEEE. 2008; **7**(4):12–18.
6. Ahlers D. Applying Geographic Information Retrieval. Datenbank-Spektrum. 2014; 14(1):39–46.
7. Goodchild MF, Yuan M, Cova TJ. Towards a general theory of geographic representation in GIS. International Journal of Geographical Information Science. 2007; 21(3):239–60.
8. Turney PD, Pantel P. From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research. 2010; 37(1):141–88.
9. Zesch T, Gurevych I. Wisdom of crowds versus wisdom of linguists-measuring the semantic relatedness of words. Natural Language Engineering. 2010; 16(01):25–59.
10. Rasheed N, et al. Semantic Representation of Abstract Words in Cognitive Robotic Model by Using Transitive Inference. Indian Journal of Science and Technology. 2014; **7**(12):2124–32.
11. Ballatore A, Wilson DC, Bertolotto M. Computing the semantic similarity of geographic terms using volunteered lexical definitions. International Journal of Geographical Information Science. 2013; 27(10):2099–118.
12. Karthikeyan K, Karthikeyani V. Ontology Based Concept Hierarchy Extraction of Web Data. Indian Journal of Science and Technology. 2015; 8(6):536–47.
13. Schwering A. Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey. Transactions in GIS. 2008; 12(1):5–29.
14. Budanitsky A, Hirst G. Evaluating Word Net-based Measures of Lexical Semantic Relatedness. Comput. Linguist. 2006; 32(1):13–47.
15. Harispe S, et al. Semantic Measures for the Comparison of Units of Language. Concepts or Entities from Text and Knowledge Base Analysis. ArXiv e-prints; 2013.
16. Wu Z, Palmer M. Verbs semantics and lexical selection in Proceedings of the 32nd annual meeting on Association for Computational Linguistics; 1994.
17. Leacock C, Chodorow M. Combining local context and Word Net similarity for word sense identification. Word Net: An electronic lexical database. 1998; 49(2):265–83.
18. Hirst G, St-Onge D. Lexical chains as representations of context for the detection and correction of malapropisms. Word Net: An electronic lexical database. 1998; 305:305–32.
19. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. arXiv preprint cmp-lg/9511007; 1995.
20. Lin D. An Information-Theoretic Definition of Similarity. in Proceedings of the Fifteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc; 1998.
21. Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008; 1997.
22. Banerjee S, Pedersen T. An adapted Lesk algorithm for word sense disambiguation using Word Net, in Computational linguistics and intelligent text processing. 2002. p. 136–45.
23. Patwardhan S, Pedersen T. Using Word Net-based context vectors to estimate the semantic relatedness of concepts. in EACL Workshop Making Sense of Sense --- Bringing Computational Linguistics and Psycholinguistics TogetherWorkshop Making Sense of Sense---Bringing Computational Linguistics and Psycholinguistics Together; 2006.
24. Ponzetto SP, Strube M. Knowledge derived from wikipedia for computing semantic relatedness. J Artif Int Res. 2007; 30(1):181–212.
25. Yeh E, et al. Wiki Walk: Random walks on Wikipedia for semantic relatedness in Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics; 2009.
26. Gouws S, Van Rooyen GJ, Engelbrecht HA. Measuring Conceptual Similarity by Spreading Activation over Wikipedia's Hyperlink Structure in Proceedings of the {2nd} Workshop on {The People's Web Meets NLP: Collaboratively Constructed Semantic Resources}. Beijing, China: Coling 2010 Organizing Committee; 2010.
27. Andrea Rodriguez M, Egenhofer MJ. Comparing geospatial entity classes: An asymmetric and context-dependent similarity measure. International Journal of Geographical Information Science. 2004; 18(3):229–56.
28. Lin Z, Lyu MR, King I. MatchSim: A novel similarity measure based on maximum neighborhood matching. Knowledge and information systems. 2012; 32(1):141–66.
29. Ballatore A, Bertolotto M, Wilson D. Geographic knowledge extraction and semantic similarity in Open Street Map. Knowledge and Information Systems. 2013; 37(1):61–81.
30. Zhang Z, Gentile AL, Ciravegna F. Harnessing different knowledge sources to measure semantic relatedness under a uniform model in Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2011.
31. Janowicz K, Raubal M, Kuhn W. The semantics of similarity in geographic information retrieval. J Spatial Information Science. 2011; 2(1):29–57.
32. Hecht B, et al. Explanatory semantic relatedness and explicit spatialization for exploratory search in Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM; 2012.

33. Dumais ST. Latent semantic analysis. Annual Review of Information Science and Technology. 2004; 38(1):188–230.
34. Stevens K, et al. Exploring Topic Coherence over Many Models and Many Topics in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg, PA, USA: Association for Computational Linguistics; 2012.
35. Lintean MC, et al. The Role of Local and Global Weighting in Assessing the Semantic Similarity of Texts Using Latent Semantic Analysis in FLAIRS Conference; 2010.
36. Gabrilovich E, Markovitch S. Wikipedia-based Semantic Interpretation for Natural Language Processing. Journal of Artificial Intelligence Research. 2009; 34:443–98.
37. Wang W, Chen P, Liu B. A Self-Adaptive Explicit Semantic Analysis Method for Computing Semantic Relatedness Using Wikipedia in Future Information Technology and Management Engineering. FITME '08. International Seminar on. 2008; 2008.
38. Liberman S, Markovitch S. Compact Hierarchical Explicit Semantic Representation in Proceedings of the IJCAI 2009 Workshop on User-Contributed Knowledge and Artificial Intelligence: An Evolving Synergy (WikiAI09). Pasadena, CA; 2009.
39. Mülligann C, et al. Analyzing the spatial-semantic interaction of points of interest in volunteered geographic information, in Spatial information theory. 2011. p. 350–70.
40. Porter MF. An algorithm for suffix stripping. Program: Electronic library and information systems. 1980; 14(3):130–7.
41. Pedersen T, Patwardhan S, Michelizzi J. Word Net: Similarity: Measuring the Relatedness of Concepts. in Demonstration Papers at HLT-NAACL 2004. Stroudsburg, PA, USA: Association for Computational Linguistics; 2004.
42. Rehurek R, Sojka P. Software Framework for Topic Modelling with Large Corpora in Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. Valletta, Malta: University of Malta; 2010.
43. Milne D, Witten IH. An open-source toolkit for mining Wikipedia. Artificial Intelligence. 2013; 194:222–39.
44. Lesk M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone in Proceedings of the 5th annual international conference on Systems documentation. ACM; 1986.